RT-Super: Learning Tumor Segmentation from Longitudinal Images and Reports

Pedro R. A. S. Bassi[1,2,3], Wenxuan Li[1,2,3], Hanxue Gu[4], Jieneng Chen[1], Xinze Zhou[1], Zheren Zhu[4], Sezgin Er[5], Ibrahim E. Hamamci[5], Bjoern H. Menze[5], Gulhan E. Akan[6], Kang Wang[4], Yang Yang[4], Alan L. Yuille[1], and Zongwei Zhou[1,7⋆]

[1] Johns Hopkins University
[2] Harvard Medical School
[3] Massachusetts General Hospital
[4] University of California, San Francisco
[5] University of Zurich
[6] Istanbul Medipol University
[7] Johns Hopkins Medicine

Abstract. Multi-tumor segmentation is important for early cancer detection and allows radiologists to visualize, verify, and understand AI predictions. However, tumor segmentation masks are expensive, time-consuming, and unavailable for many tumor types in public data. Instead, hospitals have vast, readily-available data that can guide segmentation: radiology reports and longitudinal images with multiple contrast phases. We use this readily-available data to substitute for tumor masks in training AI for tumor segmentation. To this end, we propose a new architecture, RT-Super. It has a teacher network, which analyzes the patient's longitudinal images and reports to create high-quality tumor masks. These masks train a student network, which sees a single image and no report. In inference, when longitudinal images and reports are unavailable, we use the student. RT-Super uses a new CNN-Transformer architecture and novel Consistency Losses that exploit tumors location consistency across longitudinal images. We train RT-Super to segment esophagus, uterus and spleen tumors, which have few or no public masks. Even without training masks, RT-Super can segment these tumors and surpass public AI models. Overall, we demonstrate that learning from longitudinal images, multi-phase images, and reports can overcome mask scarcity and advance multi-cancer detection and segmentation. Code: https://github.com/MrGiovanni/RT-Super

Keywords: Tumor Segmentation · Longitudinal Data · Reports

1 Introduction

Early detection of multiple cancer types can greatly improve patient survival [4,6,14], but multi-tumor segmentation is challenging: tumor segmentation masks

⋆ Correspondence to: Zongwei Zhou (zzhou82@jh.edu)

are expensive and time-consuming to create, mostly unavailable in hospitals, and public datasets offer masks for only a few tumor types. For example, public computed tomography (CT) datasets mainly provide masks for tumors in the lungs, liver, pancreas, kidneys, and colon [12,16].

While tumor masks are scarce, hospitals routinely collect rich alternative data that can guide tumor segmentation: (1) *Longitudinal images*: past and future scans of the same patient, where tumors that were small and subtle in past scans may become larger clearer in future scans; (2) *Multi-phase images*: CT scans with different contrast phases, where intra-venous contrast can improve tumor visibility; and (3) *Radiology reports*: expert-written text often describing tumor size, location, and count. These data sources are readily available in hospitals and can effectively substitute for or augment scarce tumor masks. Thus, we ask: *can AI architectures and training methods exploit reports, longitudinal images, and multi-phase images to learn tumor segmentation with fewer or no tumor masks?*

We hypothesize that leveraging longitudinal images, multi-phase images, and radiology reports in training enables AI to segment multiple tumor types without abundant segmentation masks and improves inference performance with only a single image and no report. This hypothesis stems from our observation that, when asked to segment tumors, radiologists often request all available patient images across time, multi-phase images, and all their reports. By looking at future scans or contrast-enhanced scans, radiologists can better find and segment tumors that are small or barely visible in previous or non-contrast scans.

Inspired by this, we propose the RT-Super architecture (Report & Time Supervision). During training, a teacher network receives as input longitudinal and multi-phase images and all their reports. By exploring this comprehensive information, the teacher can create higher quality tumor masks. In the absence of ground-truth tumor masks, the masks made by the teacher are used to train a student network, which only receives one image and no report. Therefore, the student can be used at inference when longitudinal/multi-phase images and reports are unavailable[1]. While the student learns from the teacher masks, the teacher learns from Report Supervision Losses [1], and a new Consistency Loss. The new loss exploits consistencies across longitudinal images, so that images where tumors are larger help the teacher segment images where tumors are smaller. For efficiency, the student is part of the teacher network (self-distillation). With RT-Super, we demonstrate that learning from reports, longitudinal, and multi-phase images improves tumor segmentation even when a single image and no report is available at inference.

To train RT-Super, we used a large-scale dataset, with 34,152 CT scans. This dataset focuses on esophagus, spleen and uterus tumors, because these tumor types have scarce public tumor masks (less than 60 for uterus and esophagus [8], 0 for spleen). Our dataset exemplifies what is readily available in hospitals: reports, longitudinal images, and multi-phase images, but no masks. Here, we train both

[1] The first diagnostic image often miss prior images; reports are always assumed missing, as AI does not need to detect tumors already detected and described in reports.

with many reports and no tumor mask, and with many reports and fewer tumor masks (767, created by our collaborating radiologists). In tumor detection, RT-Super substantially surpassed public AI models such as Google's MedGemma [18], Stanford's Merlin [5], and Universal Lesion Segmentation (ULS) [11] (Tab. 1). It also surpassed alternative tumor segmentation methods trained in our dataset, such as CLIP [17,5], multi-task learning (MTL) [21], and self-supervised learning (Models Genesis) [22] (Tab. 1). Our main contributions are:

1. RT-Super an architecture that learns multi-tumor segmentation with fewer or no masks. It learns from radiology reports, longitudinal, and multi-phase images. However, it improves performance even with a single image and no report at inference (Tab. 1, 2).
2. We created the Consistency Loss, which explores the consistency across longitudinal and multi-phase images to improve tumor segmentation.
3. We release RT-Super, which largely surpassed public AI models in the detection and segmentation of esophagus, spleen, and uterus tumors (Tab. 1).

Related work. Previous work explored radiology reports, readily available in hospitals, for learning tumor segmentation. Multi-task learning strategies extract classification labels (e.g., tumor presence or absence) from radiology reports. These labels train the AI model for classification, while the available tumor masks train it for segmentation [21]. Foundational vision-language models were trained on reports with contrastive losses (e.g., CLIP [17]), and fine-tuned for segmentation with masks [5]. These strategies used reports to learn auxiliary tasks (classification or CLIP). Instead, Report Supervision [1] used reports to directly supervise segmentation, introducing loss functions that enforce segmented tumors to match tumor count, sizes, and locations described in reports. This approach substantially improved performance. However, prior work has not used longitudinal and multi-phase CT scans as an additional source of information to further substitute for masks in learning tumor segmentation.

2 RT-Super

RT-Super is based on a new self-distillation architecture. The student is a standard segmentation model, of any architecture (here, MedFormer [10]). It receives a single image (no report) and segments tumors. The teacher is a hybrid between transformers and convolutional networks. It receives information from radiology reports, longitudinal images, and multi-phase images. The teacher exploits this information to refine the deep features of the student, creating an improved tumor segmentation mask. This mask is used to train the student in the absence of ground-truth tumor masks. The RT-Super architecture is shown in Fig. 1, and explained in Sec. 2.1. To better explore consistencies across multi-phase and longitudinal images, we introduce Consistency Losses. Before training RT-Super, we use a large-language model (LLM) to extract relevant tumor information from reports, including tumor number, locations (organs), and diameters [1]. The LLM runs only once. We use Llama 3.1 70B AWQ LLM, shown to extract tumor information from reports with 96% accuracy [3].

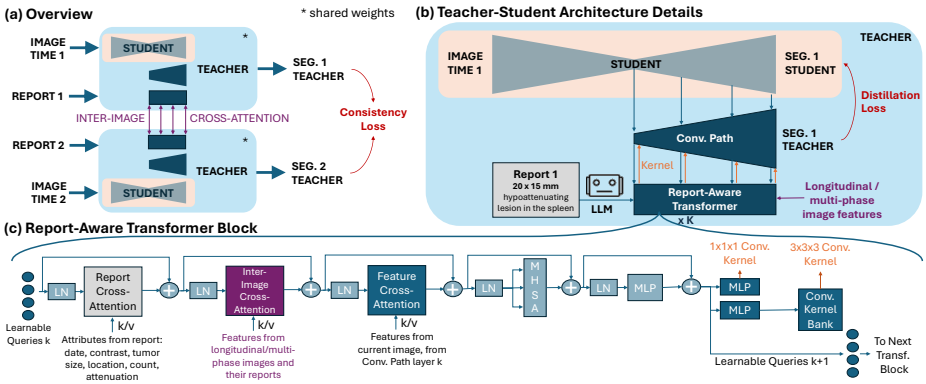

Fig. 1. Overview of RT-Super. (a) Multi-image training. RT-Super is a teacher-student framework. For a patient with multiple scans (e.g., different time points or contrast phases), we run one teacher-student instance (same weights) per scan. Each student process one scan independently. The teachers exchange information via inter-image cross-attention and are trained with an additional consistency loss across scans. **(b) Teacher-student distillation.** The teacher leverages information from multiple images and from reports to create high-quality tumor segmentation masks. The student sees one image only, but it is trained with the high-quality masks created by the teacher. At inference, when multiple images and reports are often *not* available, we use only the student. **(c) Report-aware transformer.** The teacher has a transformer that receives information from longitudinal images and reports via cross-attention. It uses this information to create convolutional layers, which refine features from the student to create an improved, report- and longitudinal-aware tumor mask. As the teacher creates this mask by refining the student's features, the student is *part* of the teacher.

2.1 RT-Super Architecture

RT-Super uses a self-distillation, teacher-student architecture, where the student is the initial part of the teacher network (Fig. 1). Any segmentation architecture can be used as the student (we used MedFormer [10]). The student receives one image and creates its tumor mask. The teacher improves this mask, by leveraging information from reports and from longitudinal images. The teacher has a *convolutional path* and a *report-aware transformer*. The convolutional path refines the deep features of the student with a sequence of convolutional layers. The kernels of these layers are created by the report-aware transformer, which has access to the student's deep features, and to information from radiology reports and multiple images. Thus, the transformer creates convolutional kernels that iteratively refine the student's features, finally creating a tumor mask that better matches reports and is more consistent with longitudinal and multi-phase images. This mask is the teacher's output.

The teacher's convolutional path is a sequence of convolutional blocks, one for each resolution level in the student decoder. The teacher's convolutional block k refines the student decoder features at level k, S_k. Specifically, the input of the teacher convolutional block k is S_k, concatenated with T_{k-1}, the outputs

of the teacher's previous convolutional block (up-sampled). Each convolutional block applies a $3 \times 3 \times 3$ convolution followed by a $1 \times 1 \times 1$ convolution (each with instance normalization and leaky ReLU). Crucially, the kernels of these two convolutions are produced by the teacher's report-aware transformer. After the teacher's final convolutional block, a $1 \times 1 \times 1$ convolution (transformer-generated) maps the final refined features to a tumor segmentation map (teacher's output).

The teacher's report-aware transformer (Fig. 1) starts from learnable queries. These queries are updated by transformer blocks, encoding information from reports and longitudinal/multi-phase images (explained later). Then, the updated queries are used to create convolutional kernels. In transformer block k, a *report cross-attention layer* updates the queries with cross-attention to tumor attributes extracted from reports (keys/values)[2]. Afterwards, a *feature cross-attention layer* updates the queries with cross-attention to the input of convolutional block k (which includes deep features from the student), followed by standard multi-head self-attention and an MLP. Updated queries at the output of transformer block k are used to create kernels for the convolutional block k. For the $1 \times 1 \times 1$ convolution, an MLP directly creates the convolutional kernel from multiple queries. For the $3 \times 3 \times 3$ convolution, directly generating its kernel would be expensive. So, we use an approach inspired by soft mixture of experts and Dynamic Convolutions [9]: the transformer chooses which $3 \times 3 \times 3$ kernel to use, from a learnable bank of M candidate kernels (we use M=16). An MLP with softmax activation takes one query from the output of the transformer block k, and predicts mixture weights for this bank. The final $3 \times 3 \times 3$ kernel is then computed as the weighted linear combination of the M kernels in the bank.

The transformer blocks are connected sequentially. Each block k updates its input queries. Some of its output queries are used to create the convolutional kernels for the convolutional block k, the others are given to the next transformer block, k+1. A special transformer block is placed before all kernel-generating blocks. It updates all learnable queries and forwards them to the subsequent blocks. In this update, the block performs cross-attention between the queries and 3 concatenated segmentation masks: the student's output, a mask for the organ with tumors, and a mask for the tumor slices. The tumor slices and the organs with tumors are defined by the report, and the corresponding organ masks are created before training[3]. These masks spatially inform the report-aware transformer where the student thinks the tumor is, and where the tumor should actually be, according to the report.

When longitudinal or multi-phase images are available for a patient, we run one teacher-student instance per image (shared weights). Each student instance processes only its own image. In contrast, the teacher instances communicate through cross-attention. Specifically, in each teacher transformer block

[2] Attributes: tumor diameter, organ, slice (tumor z coordinate), attenuation (hyper-/hypo-/iso-attenuating), and malignancy. Attributes are encoded numerically.

[3] We used nnU-Net [13] trained for organ segmentation in AbdomenAtlas [15]. Public AI [20,2] can create masks for multiple organs, but not their tumors. We use 2 cm of binary dilation to compensate for mask errors.

k, we add an *inter-image cross-attention* layer after the report cross-attention (Fig. 1). This layer updates the current instance's learnable queries by attending to the *other* teacher instances—to the input features of the convolutional block k in each other teacher instance.

We also inform the teacher about the date for each image and its contrast phase, with new tokens in the report cross-attention. Thus, when generating convolutional kernels, each teacher instance can leverage reports together with information from longitudinal and multi-phase images. RT-Super handles missing data. When report attributes or longitudinal data are missing, we set them to zero in report cross-attention. As RT-Super targets both tumor segmentation and detection, we add a classifier on top of the teacher and student segmentation outputs, it is trained jointly with the segmenter.

2.2 RT-Super Loss Functions

For images that have ground-truth tumors masks, we directly use these masks to train the teacher and the student, using the usual Dice loss and binary cross entropy losses. For images without masks, we train the teacher with Report Supervision losses [4]. The Volume Loss teaches the teacher to segment tumors matching the tumor volume and locations (organs) estimated from reports. The Ball Loss enforces the segmented tumors to match reports in terms of tumor diameter, quantity, and locations.

For images without masks, we train the student with distillation: Dice and Binary Cross-Entropy (BCE) losses encourage the student's output to match the teacher's output. We do distillation with hard targets, created by binarizing the teacher's tumor segmentation output with report-based post-processing[4] [1]. Through distillation, a student that sees a single image and no report can learn from high-quality masks created by a teacher that exploits privileged information—longitudinal images, multi-phase images, and their reports.

We also propose a **Consistency Loss** that lets the teacher use longitudinal scans with larger, clearer tumors to guide segmentation in scans where tumors are smaller and unclear. As tumors can grow over time or shrink with treatment, we enforce *location consistency*: tumors segmented in a "small-tumor" scan S should lie inside the corresponding tumors in a "large-tumor" scan L (plus a conservative margin). We apply the Consistency Loss when the radiology report indicates that all tumors in scan S are smaller than all tumors in scan L. **(1) Registration.** We register scan L to S using UniGradIcon [19], which predicts a deformation field based on images and organ masks [5]. With the deformation field, we register the teacher's output tumor mask, made for scan L, to the space of S. The teacher's output tumor mask is binarized with report-based post-processing before registration, improving the agreement between the tumor

[4] Ball Loss pseudo-mask [1]. This post-processing refines a soft mask into a binary mask that better matches the tumor size, quantity and location in reports [1].

[5] We fine-tune UniGradIcon during segmentation training. If registration fails (<0.8 DSC between registered and target organ masks), we skip the Consistency Loss.

Table 1. Training with longitudinal and multi-phase images and reports, RT-Super improves multi-tumor detection with a single image and no report at inference (internal eval). RT-Super surpasses previous public AI models and alternative training methods. The main gain over R-Super is when training without masks. This dataset includes malignant tumors and control (no-tumor) cases. Sensitivity, spec., and F1 at the operating point maximizing balanced accuracy for each model.

	train			spleen			esophagus			uterus			average		
model	longi.	report	mask	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
public AI models															
Merlin [5]		x		0	100	0	0	100	0	0	100	0	0	100	0
ULS [11]			x	28	89	**34**	5	98	**10**	32	85	**42**	22	91	**29**
MedGemma [18]		x		6	95	10	0	100	0	0	100	0	2	98	3
trained on our dataset — report only															
classification†		x		24	78	24	78	85	84	75	82	75	59	82	61
R-Super† [1]		x		88	87	72	71	88	80	92	82	82	84	86	78
RT-Super†	x	x		71	93	**73**	89	80	**90**	87	88	**84**	82	87	**82**
trained on our dataset — report and/or mask															
CLIP [5]		x	x	77	81	63	82	85	87	87	80	81	82	82	77
Multi-task l. [7]		x	x	77	83	66	76	94	85	92	83	**85**	82	87	79
Models G. [22]			x	79	74	58	83	85	87	86	78	79	83	79	75
segmentation [10]			x	91	69	62	82	90	88	88	80	82	87	80	77
nnU-Net			x	68	68	50	79	88	86	80	88	82	76	81	73
R-Super [1]		x	x	87	90	78	86	94	**91**	86	85	83	86	90	84
RT-Super	x	x	x	83	95	**83**	88	89	**91**	84	82	80	85	89	**85**

mask and the report. We dilate the registered tumor mask by 2 cm to compensate for registration errors. **(2) Consistency.** We enforce that tumors segmented in scan S lie inside the scan L registered and dilated tumor mask, by modifying the Volume Loss and Ball Loss [1]. Originally, these losses encourage segment tumors to lie inside the organ where the report mentions tumors, localized with a pre-saved organ mask. Here, we replace that organ mask with its intersection with the registered (and dilated) tumor mask from L, and then apply the Ball and Volume losses on S.

3 Results

Datasets. (1) Training: 34,152 CT-Report pairs, from 10,385 control patients (10.4% with longitudinal CTs, 3.8% multi-phase), 7,597 spleen tumor patients (33.2% longitudinal, 11.9% multi-phase), 4,258 uterus tumor patients (23.7% longitudinal, 8.8% multi-phase), and 741 esophagus tumor patients (28.7% longitudinal, 21.3% multi-phase). Includes 767 tumor masks (238 spleen, 288 uterus, 241 esophagus). All data was collected form the UCSF hospital and affiliated institutions across California. Longitudinal and multi-phase CT scans are parts of the full dataset, and all models were trained with the full dataset. **(2) Internal Test (pathology proven):** CT-Report pairs from patients with pathology-proven malignant tumors from UCSF, plus random controls—148 no-tumor, 34 spleen tumor CTs, 235 esophagus, and 84 uterus. **(3) External Test (small tumors):** CT-Report pairs from an unseen hospital (Medipol, Istanbul, Turkey). Positive scans: spleen 135, esophagus 45, uterus 12; 226 controls. Includes benign and malignant tumors, all small ($\leq$ 2 cm in diameter). **(4) External Test (masks):** used for DSC calculation, CT-Mask pairs from same external hospital,

Table 2. External validation on small tumors and ablations. We test RT-Super on an unseen hospital. Trained with reports only, RT-Super substantially surpasses R-Super in segmentation DSC. Detection (Se, Sp, AUC) is evaluated on small tumors (<2 cm) only; DSC on masks (35% small). DSC is lower for these tumor types than for others such as pancreatic tumors [16], as they are harder to segment on CT, which is not their primary imaging modality. For this difficulty, the DSC between masks created by two radiologists (2 and 8 years of experience) in our test set were 66, 50, 62 for spleen, esophagus and uterus tumors, respectively. Sensitivity and specificity at the operating point maximizing balanced accuracy for each model.

	train			spleen				esophagus				uterus				average			
model	longi.	report	mask	Se	Sp	AUC	DSC	Se	Sp	AUC	DSC	Se	Sp	AUC	DSC	Se	Sp	AUC	DSC
public AI models																			
Merlin [5]		x		1	99	50	-	0	100	50	-	33	87	60	-	11	95	53	-
ULS [11]			x	58	51	**53**	**4**	59	87	**74**	**11**	50	74	**66**	**12**	56	71	**64**	**9**
MedGemma [18]		x		0	100	50	-	0	99	50	-	0	100	50	-	0	100	50	-
trained on our dataset — report only																			
classification†		x		22	85	54	-	73	72	72	-	75	56	72	-	57	71	66	-
R-Super† [1]		x		80	85	**85**	27	68	90	82	5	58	83	78	15	69	86	**82**	16
RT-Super†	x	x		67	91	81	**31**	83	84	**86**	**33**	30	90	**79**	**25**	60	88	**82**	**30**
trained on our dataset — report and/or mask																			
nnU-Net			x	60	80	71	26	62	90	78	17	25	89	58	43	49	86	69	29
R-Super [1]		x	x	73	82	81	**54**	89	80	90	18	83	69	**80**	51	82	77	**84**	**41**
RT-Super	x	x	x	66	89	**82**	43	89	82	**91**	**24**	75	70	78	**55**	77	80	**84**	**41**
ablation of RT-Super — report and/or mask																			
no consistency loss	x	x	x	62	93	82	44	86	77	89	21	58	78	78	52	69	83	83	39
no inter-image cross-att	x	x	x	71	83	81	21	82	86	91	16	70	74	80	53	74	81	84	30
no dynamic kernel	x	x	x	73	91	83	47	84	87	89	24	75	72	79	48	77	83	84	40
no size-info teacher	x	x	x	69	86	80	45	71	87	86	19	100	59	80	42	80	77	82	35
no report-info teacher	x	x	x	65	87	83	37	86	85	90	18	83	70	77	52	78	81	83	36

spleen 25, esophagus 17, uterus 13. Benign and malignant tumors, 35% small. In testing, AI models have no access to patient reports or to longitudinal images.

Training RT-Super: We trained two variants: RT-Super† using CT–Report pairs, and RT-Super using CT–Report and CT-Mask pairs. We used MedFormer [10] as the student, and used its training hyper-parameters for RT-Super and all MedFormer-based baselines—segmentation, CLIP, ModelsGenesis, MTL, R-Super, and classification (MedFormer's encoder). We used the R-Super cropping strategy [1]. Loss weights were 1 for supervised segmentation losses and distillation, and 0.1 for Consistency and Report Supervision losses. By the end of training, the RT-Super teacher surpassed its student by +2.9% DSC. nnU-Net [13] was trained out-of-the-box (ResEncL architecture), but with 1mm spacing. ModelsGenesis / CLIP pre-trained on all CTs / CT-Report pairs, and fine-tuned on all CT-Mask pairs. MTL trained jointly on all CT-Report and CT-Mask pairs. Classification trained on all CTs, with tumor presence/absence labels (per organ) from reports. Segmentation (MedFormer) and nnU-Net were only trained on CT-Mask pairs. ULS, Merlin, and MedGemma used their public checkpoints.

RT-Super surpasses prior public AI in detecting esophagus, spleen and uterus tumors (Tab. 1 and 2). RT-Super substantially surpasses Merlin, MedGemma, and ULS, leading public AI models. Merlin and MedGemma are VLMs. They generated reports and we evaluated these reports for tumor detection (following [3]). The VLM reports had low sensitivity, missing the tumor types we analyzed here. VLMs were surpassed by segmentation models, as

in [3]. Here, VLM performances were even lower than in [3], because our tumor types are rarer and more difficult to detect on CT. ULS surpassed VLMs, but still had low sensitivity, likely because, if present, our tumor types are underrepresented in its training data. Also, ULS is trained on tumor crops only, causing a disadvantage for tumor detection. RT-Super surpassed all the public models.

RT-Super improves tumor detection & segmentation with few or no tumor masks. Tab. 1 shows that RT-Super surpassed 8 other training methods that also trained on our dataset, but were not designed to jointly leverage reports, longitudinal, and multi-phase images. Although RT-Super performed similarly to R-Super in tumor detection, it substantially surpassed R-Super in segmentation DSC when training without masks (Tab. 2). Tab. 2 displays diverse ablation studies, showing how each component of RT-Super improves performance.

Conclusion. Tumor mask creation is expensive and time-consuming, leaving public datasets and segmentation models unable to cover most tumor types. Following R-Super [1,4], RT-Super shows that AI can learn tumor segmentation from routinely available hospital data: reports, longitudinal data, and multi-phase images, rather than manual masks, scaling tumor segmentation.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the National Institutes of Health (NIH) under Award Number R01EB037669. Paper content is covered by patents pending.

Disclosure of Interests. The authors declare no competing interests.

References

1. Bassi, P.R., Li, W., Chen, J., Zhu, Z., Lin, T., Decherchi, S., Cavalli, A., Wang, K., Yang, Y., Yuille, A.L., Zhou, Z.: Learning segmentation from radiology reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 305–315. Springer (2025), https://github.com/MrGiovanni/R-Super
2. Bassi, P.R., Li, W., Tang, Y., Isensee, F., Wang, Z., Chen, J., Chou, Y.C., Kirchhoff, Y., Rokuss, M., Huang, Z., Ye, J., He, J., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K.H., Jaeger, P., Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y., Xing, Z., Zhu, L., Sadegheih, Y., Bozorgpour, A., Kumari, P., Azad, R., Merhof, D., Shi, P., Ma, T., Du, Y., Bai, F., Huang, T., Zhao, B., Wang, H., Li, X., Gu, H., Dong, H., Yang, J., Mazurowski, M.A., Gupta, S., Wu, L., Zhuang, J., Chen, H., Roth, H., Xu, D., Blaschko, M.B., Decherchi, S., Cavalli, A., Yuille, A.L., Zhou, Z.: Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track **37**, 15184–15201 (2024), https://github.com/MrGiovanni/Touchstone
3. Bassi, P.R., Yavuz, M.C., Hamamci, I.E., Er, S., Chen, X., Li, W., Menze, B., Decherchi, S., Cavalli, A., Wang, K., Yang, Y., Yuille, A., Zhou, Z.: Radgpt: Constructing 3d image-text tumor datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23720–23730 (2025), https://github.com/MrGiovanni/RadGPT

4. Bassi, P.R., Zhou, X., Li, W., Płotka, S., Chen, J., Chen, Q., Zhu, Z., Prządo, J., Hamamci, I.E., Er, S., Chen, X., Yavuz, M.C., Chou, Y.C., Lin, T., Wang, K., Tang, Y., Cwikla, J.B., Decherchi, S., Cavalli, A., Yang, Y., Yuille, A.L., Zhou, Z.: Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks. arXiv preprint arXiv:2510.14803 (2025), https://github.com/MrGiovanni/R-Super
5. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square pp. rs–3 (2024)
6. Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., et al.: Large-scale pancreatic cancer detection via non-contrast ct and deep learning. Nature medicine **29**(12), 3033–3043 (2023)
7. Chen, E.Z., Dong, X., Li, X., Jiang, H., Rong, R., Wu, J.: Lesion attributes segmentation for melanoma detection with multi-task u-net. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 485–488. IEEE (2019)
8. Chen, Q., Zhou, X., Liu, C., Chen, H., Li, W., Jiang, Z., Huang, Z., Zhao, Y., Yu, D., He, J., Zheng, Y., Shao, L., Yuille, A., Zhou, Z.: Scaling tumor segmentation: Best lessons from real and synthetic data. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 24001–24013 (2025), https://github.com/BodyMaps/AbdomenAtlas2.0
9. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11030–11039 (2020)
10. Gao, Y., Zhou, M., Liu, D., Yan, Z., Zhang, S., Metaxas, D.N.: A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131 (2022)
11. de Grauw, M., Scholten, E.T., Smit, E.J., Rutten, M.J., Prokop, M., van Ginneken, B., Hering, A.: The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. Medical Image Analysis **102**, 103525 (2025)
12. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
13. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
14. Li, W., Bassi, P.R.A.S., Wu, L., Zhou, X., Zhao, Y., Chen, Q., Plotka, S., Lin, T., Zhu, Z., Martin, M., Caskey, J., Jiang, S., Chen, X., Ćwikła, J.B., Sankowski, A., Wu, Y., Decherchi, S., Cavalli, A., Lall, C., Tomasetti, C., Guo, Y., Yu, X., Cai, Y., Qiao, H., Bao, J., Hu, C., Wang, X., Sitek, A., Ding, K., Li, H., Wang, M., Yu, D., Zhang, G., Yang, Y., Wang, K., Yuille, A.L., Zhou, Z.: Early and prediagnostic detection of pancreatic cancer from computed tomography. arXiv preprint arXiv:2601.22134 (2026), https://github.com/BodyMaps/ePAI
15. Li, W., Qu, C., Chen, X., Bassi, P.R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., Tang, Y., Cao, Y., Han, H., Zhang, Z., Liu, J., Zhang, T., Ma, Y., Wang, J., Zhang, G., Yuille, A., Zhou, Z.: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis p. 103285 (2024), https://github.com/MrGiovanni/AbdomenAtlas

16. Li, W., Zhou, X., Chen, Q., Lin, T., Bassi, P.R., Chen, X., Ye, C., Zhu, Z., Ding, K., Li, H., Wang, K., Yang, Y., Tang, Y., Xu, D., Yuille, A.L., Zhou, Z.: Pants: The pancreatic tumor segmentation dataset. In: Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (2025), https://github.com/MrGiovanni/PanTS
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
18. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
19. Tian, L., Greer, H., Kwitt, R., Vialard, F.X., San José Estépar, R., Bouix, S., Rushmore, R., Niethammer, M.: unigradicon: A foundation model for medical image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 749–760. Springer (2024)
20. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
21. Zhang, Y., Li, H., Du, J., Qin, J., Wang, T., Chen, Y., Liu, B., Gao, W., Ma, G., Lei, B.: 3d multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. IEEE transactions on medical imaging **40**(6), 1618–1631 (2021)
22. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical Image Analysis **67**, 101840 (2021), https://github.com/MrGiovanni/ModelsGenesis